Régression inverse par tranches pour une population stratifiée

Marie Chavent[1,2] & Vanessa Kuentz[3] & Benoît Liquet[4] & Jérôme Saracco[1,2]

[1] *Institut de Mathématiques de Bordeaux, UMR CNRS 5251*
Université de Bordeaux
351 cours de la libération, 33405 Talence Cedex, France
e-mail: {marie.chavent,jerome.saracco}@math.u-bordeaux1.fr

[2] *INRIA Bordeaux Sud-Ouest, CQFD team, France*

[3] *CEMAGREF, 50 avenue de Verdun Gazinet, 33612 Cestas cedex*
e-mail: vanessa.kuentz@bordeaux.cemagref.fr

[4] *INSERM U897, ISPED*
Université Victor Segalen Bordeaux 2
146 rue Leo Saignat, 33076 Bordeaux Cedex, France
e-mail: benoit.liquet@isped.u-bordeaux2.fr

Résumé. Dans cette communication, nous considérons un modèle semiparamétrique de régression dans lequel une variable à expliquer Y dépend d'une covariable quantitative X de dimension p et d'une variable qualitative Z. Cette covariable Z définit une stratification de la population. Ce modèle inclut une reduction de sa partie explicative via un indice $X'\beta$. Nous proposons une approche fondée sur la méthode SIR (Sliced Inverse Regression ou régression inverse par tranches en français) afin d'estimer la direction du vecteur de paramètre β. Nous avons obtenu des résultats asymptotiques pour l'estimateur proposé (convergence et normalité asymptotique). Des simulations ont montré le bon comportement numérique de l'estimateur dans les cas homoscédastique et hétéroscédastique.
Mots-clés : réduction de dimension, sliced inverse regression (SIR), variable qualitative, décomposition aux valeurs propres.

Abstract. In this communication, we consider a semiparametric single index regression model involving a real dependent variable Y, a p-dimensional quantitative covariable X and a categorical predictor Z which defines a stratification of the population. This model includes a dimension reduction of X via an index $X'\beta$. We propose an approach based on sliced inverse regression in order to estimate the space spanned by the common dimension reduction direction β. We establish $\sqrt{n}$-consistency of the proposed estimator and its asymptotic normality. Simulation study shows good numerical performance of the proposed estimator in homoscedastic and heteroscedastic cases. Extensions to multiple indices models, q-dimensional response variable and/or SIR_α based methods are also possible. The case of unbalanced subpopulations is treated. Finally a practical method to investigate if there is or not a common direction β is proposed.
Keywords: dimension reduction, sliced inverse regression (SIR), categorical covariate, eigen decomposition.

1 Introduction

Soit Y une variable à expliquer et soit X un variable explicative de dimension p. Une covariable qualitative Z à L modalités est aussi disponible et engendre une stratification de la population en L sous-populations. Nous considérons dans cette communication le modèle semiparamétrique de régression suivant : pour $l = 1, \ldots, L$,

$$Y = g^{(l)}(X'\beta, \varepsilon) \quad \text{quand} \quad Z = l, \tag{1}$$

où ε est un terme d'erreur aléatoire indépendant de X. Pour chaque sous-population l, Y est reliée à X seulement via l'indice $X'\beta$. La covariable qualitative Z n'entre pas dans l'aspect "réduction de dimension" du modèle, elle permet seulement d'identifier les L sous-populations. On suppose de plus que la distribution conditionnelle de X sachant Z est telle que $\mathbf{E}(X|Z = l) = \mu^{(l)}$ et $\mathbf{V}(X|Z = l) = \Sigma^{(l)}$ existent pour $l = 1, \ldots, L$. Notons que la covariable Z a un effet sur la dépendance entre Y et l'indice $X'\beta$ via les fonctions de lien $g^{(l)}$ associées à chacune des sous-populations. En d'autres termes, on peut aussi voir le modèle (1) comme suit : Y et (X, Z) sont indépendants conditionnellement à $(X'\beta, Z)$. Dans ce modèle, notre objectif est d'estimer la direction du vecteur de paramètre β commun aux L sous-populations (le vecteur β n'étant pas totalement identifiable). Les fonctions de lien $g^{(l)}$ pourront ensuite être estimées non paramétriquement par des estimateurs à noyau ou de type splines de lissage par exemple. Dans la suite, on va appeler direction EDR (pour "effective dimension reduction") toute direction de $\Re^p$ colinéaire à β. Comme dans le cadre standard des approches de type SIR (sliced inverse regression ou régression inverse par tranches en français, voir par exemple Duan et Li (1991) ou Li (1991) pour une présentation de SIR), la condition de linéarité suivante est nécessaire :

(LC) Pour chaque sous-population $l = 1, \ldots, L$,

$$\mathbf{E}(X'v|X'\beta, Z = l) \text{ est linéaire en } X'\beta \text{ pour tout } v \in \Re^p.$$

Contrairement à certaines approches de type SIR devéloppées dans le cadre de ce modèle avec une covariable catégorielle, l'approche proposée à la section suivante est valable dans les cas homoscédastique ($\Sigma^{(l)} = \Sigma^*$ pour tout $l = 1, \ldots, L$) et hétéroscédastique (les matrices $\Sigma^{(l)}$ peuvent être différentes).

2 Méthode d'estimation proposée

2.1 Version sur population

L'approche considérée consiste à rechercher au sein de chaque sous-population la direction de β, puis à combiner convenablement les L directions obtenues individuellement.

Pour cela, nous proposons tout d'abord de mettre en œuvre la méthode SIR dans chaque sous-population. Pour chaque sous-population l, on définit la matrice d'intéêt,

notée $M_I^{(l)}$, utilisée dans SIR.Soit $T^{(l)}$ un tranformation monontone de Y sachant que $Z = l$. Nous avons :

$$M_I^{(l)} = \mathbf{V}(\mathbf{E}(X|T^{(l)}(Y), Z = l)).$$

Afin de pouvoir estimer facilement cette matrice, Li (1991) a proposé comme transformation $T^{(l)}$ un tranchage qui discrétise la variable Y en une nouvelle variable à $H^{(l)} > 1$ niveaux : le support de Y sachant $Z = l$ est partitionné en $H^{(l)}$ tranches fixes distinctes $s_h^{(l)}$. Ainsi la matrice $M_I^{(l)}$ peut s'écrire sous la forme :

$$M_I^{(l)} = \sum_{h=1}^{H^{(l)}} p_h^{(l)}(m_h^{(l)} - \mu^{(l)})(m_h^{(l)} - \mu^{(l)})',$$

avec $p_h^{(l)} = P(Y \in s_h^{(l)}|Z = l)$, $m_h^{(l)} = \mathbf{E}(X^{(l)}|Y \in s_h^{(l)}, Z = l)$ et $\mu^{(l)} = \mathbf{E}(X|Z = l)$. Sous la condition de linéarité (LC), le vecteur propre $b^{(l)}$ associé à la plus grande valeur propre de la matrice $(\Sigma^{(l)})^{-1}M_I^{(l)}$ est une direction EDR.

Pour combiner les L directions $b^{(l)}$, nous définissons maintenant la matrice $B = [b^{(1)}, \ldots, b^{(L)}]$. Notons par b le vecteur propre associé à la plus grande valeur propre de la matrice BB'. Le théorème ci-dessous assure que ce vecteur est une direction EDR.

Théorème 1. *Sous la condition de linéarité (LC) et le modèle* (1)*, le vecteur propre b associé à la valeur propre non nulle de BB' est colinéaire à β.*

PREUVE. Pour chaque sous-population $l = 1, \ldots, L$, $b^{(l)}$ est colinéaire β, i.e. $b^{(l)} = \alpha_l\beta$, où $\alpha_l \in \Re_+^*$. Vu que $B = [\alpha_1\beta, \ldots, \alpha_L\beta]$, on a $BB' = \sum_{l=1}^{L} \alpha_l^2\beta\beta' = \|\alpha\|^2\beta\beta'$, avec $\alpha = (\alpha_1, \ldots, \alpha_L)'$ et $\|.\|$ est la norme associée au produit scalaire usuel. Ainsi le vecteur propre b associée à la seule valeur propre non nulle (positive) de BB' est colinéaire à β, et est donc une direction EDR. □

2.2 Version sur échantillon

Considérons un échantillon i.i.d. $\mathcal{S} = \{(X_i, Y_i, Z_i),\ i = 1, \ldots, n\}$ issu du modèle (1). Pour obtenir un estimateur de la matrice $M_I^{(l)}$, l'idée de l'approche SIR est de remplacer les moments théoriques par les moments empiriques correspondants. Soit $\mathcal{S}^{(l)} = \{(Y_i, X_i'), i = 1, \ldots, n^{(l)}$ tel que $Z_i = l\}$ le sous-échantillon associé à la sous-population l, où $n^{(l)}$ est la taille du sous-échantillon $\mathcal{S}^{(l)}$. Dans chaque sous-population l, on a alors :

$$\widehat{M_I}^{(l)} = \sum_{h=1}^{H^{(l)}} \hat{p}_h^{(l)}(\hat{m}_h^{(l)} - \overline{X}^{(l)})(\hat{m}_h^{(l)} - \overline{X}^{(l)})'$$

avec $\overline{X}^{(l)} = \frac{1}{n^{(l)}} \sum_{i=1}^{n^{(l)}} X_i$, $\hat{p}_h^{(l)} = \frac{1}{n^{(l)}} \sum_{i=1}^{n^{(l)}} \mathbb{I}_{[y_i \in s_h^{(l)}]}$ et $\hat{m}_h^{(l)} = \frac{1}{n^{(l)}\hat{p}_h^{(l)}} \sum_{i=1}^{n^{(l)}} X_i^{(l)}\mathbb{I}_{[y_i \in s_h^{(l)}]}$, où la notation $\mathbb{I}$ désigne la fonction indicatrice. Le vecteur propre $\hat{b}^{(l)}$ associé à la plus grande valeur propre de $(\widehat{\Sigma}^{(l)})^{-1}\widehat{M_I}^{(l)}$ avec $\widehat{\Sigma}^{(l)} = \frac{1}{n^{(l)}} \sum_{i=1}^{n^{(l)}} (X_i^{(l)} - \overline{X}^{(l)})(X_i^{(l)} - \overline{X}^{(l)})'$,

est la direction EDR estimée dans la sous-population l. On construit ensuite la matrice $\widehat{B} = [\hat{b}^{(1)}, \ldots, \hat{b}^{(L)}]$. Le vecteur propre $\hat{b}$ associé à la plus grande valeur propre de la matrice $\widehat{B}\widehat{B}'$ est la direction EDR estimée dans le modèle (1).

2.3 Résultats asymptotique

Notons $n_h^{(l)}$ le nombre d'observations dans tranche $s_h^{(l)}$. Les hypothèses suivantes sont nécessaires pour établir les résultats de convergence ci-dessous :
(A1) Chaque sous-échantillon $\mathcal{S}^{(l)}$, $l = 1, \ldots, L$, est un échantillon d'observations indépendantes issues du modèle (1).
(A2) Pour chaque sous-population l, le support de Y est partionné en $H^{(l)}$ tranches fixes telle que $p_h^{(l)} \neq 0, h = 1, \ldots, H^{(l)}$.
(A3) Pour $l = 1, \ldots, L$ et $h = 1, \ldots, H^{(l)}$, $n_h^{(l)} \to \infty$ (et donc $n^{(l)} \to \infty$) lorsque $n \to \infty$.
Théorème 2. *Sous la condition de linéarité (LC) et sous les hypothèses* ***(A1)-(A3)****, nous avons :*

$$\hat{b} = b + O_p(n^{-1/2}) \quad \text{et} \quad \sqrt{n}(\hat{b} - b) \longrightarrow_d W \sim \mathcal{N}(0, \Gamma_W).$$

La démonstration de ce théorème est disponible dans Chavent et al. (2010).

3 Quelques simulations

Nous avons fait une étude sur simulation avec le logiciel `R` afin d'illustrer le comportement numérique de l'estimateur proposé. Nous avons comparé cet estimateur avec l'approche SIR classique (c'est à dire sans tenir compte de l'existence de sous-populations) et avec les deux estimateurs proposés par Liquet and Saracco (2007), le premier étant dédié au cadre homoscédastique et le second au cadre hétéroscédastique. Dans la suite ces différents estimateurs seront respectivement appelés "new", "SIR", "homo" et "hetero".

Dans la suite, nous nous limiterons au cas d'un modèle à un seul indice avec $L = 2$ sous-populations. Dans Chavent et al. (2010), une étude plus complète est disponible avec des modèles à plusieurs indices et un nombre L de sous-populations supérieur à 2. La mesure de qualité utilisée ici est le cosinus carré de l'angle entre l'estimation $\hat{b}$ et la vraie direction β. Plus sa valeur est proche de 1, meilleure est l'estimation.

Nous avons généré des échantillons de taille $n = 200$ (avec $n^{(1)} = n^{(2)} = 100$) à partir du modèle suivant :

$$\begin{cases} Y = (\frac{1}{8}\beta' X)^3 + \epsilon_1 & \text{quand } \ Z = 1, \\ Y = -(\beta' X)/2 + \epsilon_2 & \text{quand } \ Z = 2, \end{cases} \tag{2}$$

où $X|Z = l$ (pour $l = 1, 2$) suit une loi normale multidimensionnelle ($p = 5$) de moyenne $\mu^{(l)} = \mathbf{0}_5$ et de matrice de variance (générée aléatoirement) $\Sigma^{(l)}$. La manière de générer

les matrices $\Sigma^{(l)}$ est la suivante. Les termes d'une matrice $\Lambda^{(l)}$ de dimension $p \times p$ sont générés à partir d'une loi uniforme sur $[-2, 2]$. Puis $\Sigma^{(l)} = \Lambda^{(l)}\Lambda^{(l)\prime} + 0.5I_p$ afin de ne pas avoir de problème d'inversion de $\Sigma^{(l)}$. On se trouve donc généralement dans un cadre hétéroscédastique. Le terme d'erreur aléatoire ϵ_l est indépendant de X sachant $Z = l$ et $\epsilon_l \sim \mathcal{N}(0, 0.7^2)$ pour $l = 1, 2$. Nous avons pris $\beta = (1, 2, -1, -2, 0)'$.

Etude d'un échantillon simulé. La direction EDR a été estimée par les quatre méthodes ("SIR", "hetero", "homo" and "new"). Comme "SIR" ne permet pas d'obtenir une bonne estimation (cosinus carré de 0.40), nous donnons les listes des valeurs propres associées aux trois autres méthodes: $\lambda_{\text{homo}} = (1.70, 0.83, 0.29, 0.08, 0.01)$, $\lambda_{\text{hetero}} = (0.88, 0.06, 0.06, 0.01, 0.01)$ et $\lambda_{\text{new}} = (0.99, 0.01, 0, 0, 0)$. On observe très clairement un saut entre la première et la seconde valeur propre, ainsi on ne retient qu'une seule direction EDR. Les directions EDR estimées sont : $\hat{b}_{\text{homo}} = (0.43, -0.47, 0.48, 0.22, 0.56)$, $\hat{b}_{\text{hetero}} = (-0.34, -0.64, 0.39, 0.56, 0.08)$ et $\hat{b}_{\text{new}} = (-0.35, -0.64, 0.39, 0.56, 0.08)$. Les méthodes "hetero" et "new" donnent d'excellentes estimations avec des cosinus carrés supérieur à 0.99. Le cosinus carré pour la méthode "homo" n'est que de 0.84, ce qui était attendu vu que l'on s'est placé dans le cadre d'un modèle hétéroscédastique. Sur la Figure 1, on

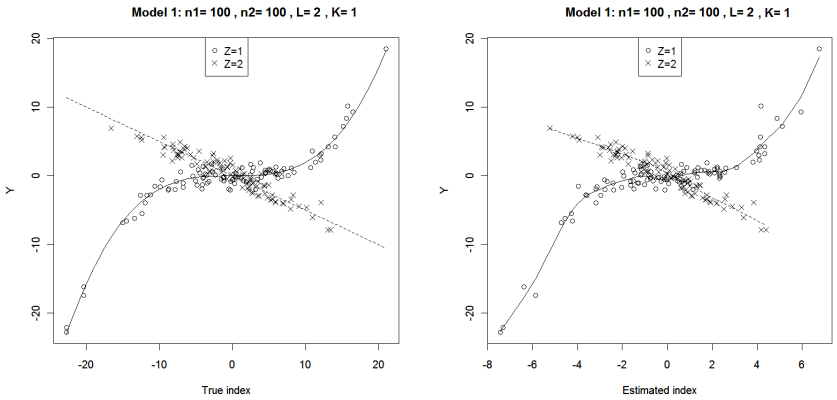

Figure 1: A gauche : nuage des vrais indices ($X'\beta$) versus Y. A droite : nuage des indices estimés ($X'\hat{b}$) versus Y avec les fonctions de lien estimées pour $Z = 1$ (ligne pleine) et pour $Z = 2$ (ligne en pointillés).

trouve à gauche (resp. à droite) les nuages, pour chaque sous-population l, de la variable réponse Y contre le vrai indice commun $X'\beta$ (resp. avec l'indice estimé $X'\hat{b}_{\text{new}}$). Sur le graphique de gauche, on a tracé aussi les vraies fonctions de liens $g^{(l)}$, alors que sur le graphique de droite, on a représenté les estimations par splines de lissage de ces fonctions.

Bilan d'une simulation. A partir du modèle (2) , $N = 500$ échantillons de taille $n = 200$ (avec $n^{(1)} = n^{(2)} = 100$) ont été simulés. Pour chaque échantillon simulé, la direction EDR a été estimé aves les méthodes "homo", "hetero", "new" et "SIR", et les cosinus carrés correspondants ont été calculés. Afin de comparer les estimateurs, on a tracé sur la Figure 2 les boxplots des ces $N = 500$ cosinus carrés. On observe que la méthode "SIR" ne parvient pas à estimer convenablement la direction EDR. Les méthodes "new" et

"hétéro" donnent d'excellentes estimations avec un léger avantage pour la méthode "new" (voir le graphique de droite à la Figure 2 qui représente le nuage croisant les cosinus carrés obtenus par la méthode "new" avec ceux obtenus par la méthode "hetero"). Notons que la méthode "new" a l'avantage de ne pas avoir de cas pathologiques comme la méthode "hétéro" (voir Liquet et Saracco (2007) ou Chavent et al. (2010) pour plus de précisions sur ce point). L'approche "homo" ne permet pas d'obtenir des bonnes estimations, ce qui met en évidence l'intérêt de considérer des approches hétéroscédastiques.

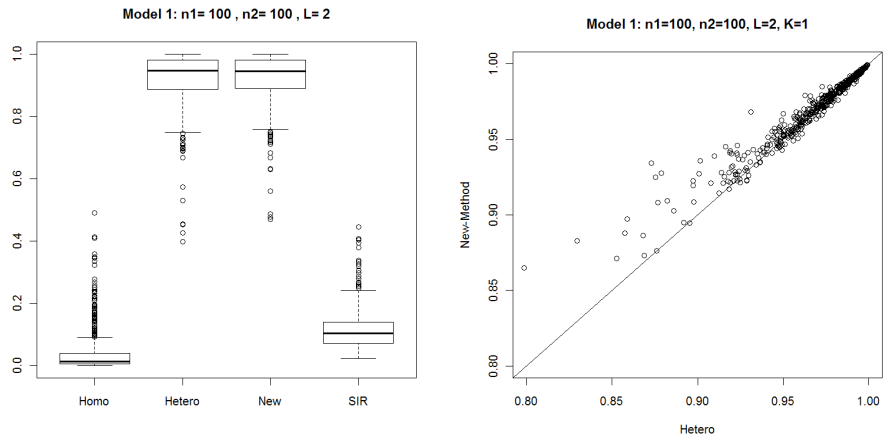

Figure 2: A gauche : boxplots des cosinus carrés obtenus par les quatre méthodes. A droite : comparaison des cosinus carrés pour les méthodes "hetero" et "new".

Remarques finales. Diverses extensions de l'approche développée sont possibles : cas de modèles à plusieurs indices $X'\beta_1 \ldots, X'\beta_K$, cas d'une variable dépendante Y de dimension q, ... Il est possible d'adapter l'estimateur au cas où les populations sont de "tailles" différentes. De plus, nous présentons aussi dans Chavent et al. (2010) une méthode pratique permettant d'examiner si une direction commune β existe bien au sein des différentes populations.

Bibliographie

[1] Chavent, M., Kuentz, V., Liquet, B. et Saracco, J. (2010). A sliced inverse regression approach for a stratified population. *Soumis pour publication.*
[2] Duan, N. et Li, K.C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, **19**, 505-530.
[3] Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-342.
[4] Liquet, B. et Saracco, J. (2007). Pooled marginal slicing approach via SIR_α with discrete covariables. *Computational Statistics*, **4**, 599-617.